FUNDAÇÃO DE ENSINO "EURÍPIDES SOARES DA ROCHA"
CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA – UNIVEM
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

DANIEL DA SILVA DISNER

MINERAÇÃO DE DADOS PARA OBTENÇÃO DE CONHECIMENTO EM BIG DATA

MARÍLIA
2014

DANIEL DA SILVA DISNER

MINERAÇÃO DE DADOS PARA OBTENÇÃO DE CONHECIMENTO EM BIG DATA

Trabalho de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Fundação de Ensino "Eurípides Soares da Rocha", mantenedora do Centro Universitário Eurípides de Marília – UNIVEM, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador
Prof.: Geraldo Pereira Junior

MARÍLIA
2014

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO – AVALIAÇÃO FINAL

Daniel da Silva Disner

Mineração de dados para obtenção de conhecimento em Big Data

Banca examinadora da monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do UNIVEM/F.E.E.S.R., para obtenção do Título de Bacharel em Sistemas de Informação.

Nota: 9,0 (nove)

Orientador: Geraldo Pereira Junior

1°. Examinador: Elvis Fusco

2°. Examinador: Jussara Mallia Zachi

Marília, 01 de dezembro de 2014.

DISNER, Daniel

Mineração de dados para obtenção de conhecimento em Big Data / Daniel da Silva Disner; Orientador: Prof. Geraldo Pereira Junior. Marília, SP: [s.n.], 2014.

41 folhas

Monografia (Bacharelado em Sistemas de Informação): Centro Universitário Eurípides de Marília.

AGRADECIMENTOS

Agradeço aos meus pais Maurino e Marlene, por toda dedicação e apoio para que eu pudesse lutar pelos meus sonhos e não desistir diante das dificuldades.

À minha irmã Gabriela, que esteve sempre ao meu lado me incentivando como fez em mais esta etapa.

Ao Professor Geraldo, por aceitar me orientar e pela amizade e ajuda em vários aspectos para eu pudesse atingir os objetivos esperados.

Por fim, agradeço a todos os professores e amigos da UNIVEM, pelos bons momentos e conhecimentos compartilhados para que eu possa me tornar um bom profissional e um ser humano melhor.

RESUMO

O grande volume de dados gerados pelos mais diversos sistemas de informações, em geral, inviabiliza a possibilidade de análise humana em um prazo razoável, sendo dificultada a identificação de relações entre dados, com padrões novos úteis e válidos sobre o comportamento ou opinião do usuário ou também que forneçam informações que possam impulsionar a melhoria de processos dentro de uma organização e fornecer a obtenção de conhecimento aplicável tanto nas áreas de pesquisas científicas ou para às áreas voltadas para a inteligência de negócios. Este trabalho teve como o principal o objetivo de avaliar e analisar a potencialidade da integração das ferramentas de código aberto WEKA e Linguagem R na construção de uma aplicação escrita na linguagem de programação JAVA para a obtenção e geração de conhecimento através da Mineração de Dados, análise estatística, representação gráfica e identificação de padrões inter-relacionados detectando relações que contribuam e auxiliem a tomada de decisões e a extração de conhecimento em grandes bases de dados. Tomando como base, dados de opiniões de usuários de redes sócias, e tendo como escolha para a extração desses dados a rede social de micro-blogging *Twitter*.

Palavras-Chave: Mineração de Dados, Extração de Conhecimento, Grandes Bases de Dados, Weka, Linguagem R, Tomada de Decisão.

ABSTRACT

The big volume of data generated by different information systems, usually prevents the possibility of human analysis in a reasonable time, to identify relationships between data being difficult with useful and valid new standards about behavior or user opinion or also that provide information that may drive improvement processes within an organization and provide the applicable attainment of knowledge both in the areas of scientific research or for the areas facing the business intelligence. This work was the principal to evaluate and analyze the potential of integrating open source tools WEKA and R Language in building an application written in the Java programming language for the collection and generation of knowledge through data mining, analysis statistical, graphical representation and identification of interrelated standards by detecting relationships that contribute and assist the decision making and the extraction of knowledge in big data. On the basis, data of members networks users reviews, and with the choice for extracting these data the social network of micro-blogging Twitter.

Keywords: Data Mining, Extraction of Knowledge, Big Data, Weka, R Language, Decision Making.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

KDD	Knowledge Discovery in Databases (Descoberta de Conhecimentos em Bancos de Dados)
API	Application Programming Interface (Interface de Programação de Aplicativos)

SUMÁRIO

INTRODUÇÃO

O uso de sistemas computacionais por pessoas ou organizações geram diariamente uma enorme quantidade de dados através de registros de ações de usuários, postagens em redes sociais, transações feitas em sistemas empresariais. Comumente muitos desses dados são ignorados, principalmente pela inviabilidade de se fazer uma análise e relacionar esses dados para tomar algum tipo de decisão, que possa agregar alguma melhoria ou conhecimento de uma forma ágil e sem a ajuda de recursos ou ferramentas associadas à tecnologia da informação.

O termo Big Data é utilizado para representar essa quantidade enorme de dados gerados diariamente e que podem dar às organizações vantagens competitivas e promover inovações, no entanto é um termo no qual não há um consenso capaz de mensurar ou definir com clareza o que realmente é Big Data, e as tecnologias fundamentais que a sustentam. [TAURION, 2013]

Para se realizar alguma espécie de análise dentro dessa enorme gama de informações em um prazo razoável, será difícil através dos métodos tradicionais pois estes métodos são capazes de tratar apenas as informações que estão explícitas, então dessa forma torna-se necessário utilizar-se de ferramentas para se efetuar um procedimento de mineração nessas grandes bases de dados (Data Mining), explorando esses dados em busca de relações e padrões consistentes, encontrando relacionamentos entre variáveis e identificando dessa forma novos subconjuntos de dados, utilizando vários recursos e técnicas da estatística, inteligência artificial e recuperação da informação, para transformar essa enorme quantidade de dados brutos em informação útil, extraindo e gerando conhecimento onde seria humanamente inviável. [NAVEGA, 2002]

No contexto de se realizar uma análise inteligente de dados em massa utilizando um recurso de *Data Mining*, utiliza-se o uso de um conceito conhecido como KDD (*Knowledge Discovery in Databases*) que significa de uma maneira simplificada: extração (ou descoberta) de conhecimentos em bases de dados que é um processo usado para a identificação de padrões anteriormente desconhecidos e que tenham valor para a organização que irá fazer uso ou implementar uma ferramenta mineradora, em análise de um certo volume de dados concluindo assim uma extração de conhecimento, podendo revelar informações importantes e relevantes que possam de alguma maneira influenciar de forma positiva a estratégia corporativa, ou até mesmo ajudar a melhoraria de processos internos e aumentar ainda a

lucratividade de uma empresa ou comércio.

Existe uma necessidade de se utilizar mecanismos para a mineração desses dados de uma forma eficiente e rápida, de forma com que esses dados possam ser aproveitados para gerar, extrair ou fazer a descoberta de conhecimentos e assim contribuir para a prevenção de situações que ofereçam risco à operações em uma empresa ou integridade do usuário, melhorar processos internos a fim de reduzir custos e aumentar a rentabilidade de um negócio, podendo assim beneficiar uma infinidade de áreas [Olson, 2008 e Witten, 2005].

As ferramentas e recursos utilizados para se fazer um processo de mineração de dados utilizam-se de uma série de princípios que levam a conclusões através de processos de indução, diferente das técnicas comuns de banco de dados que oferecem uma conclusão dedutiva, que são inadequadas para se encontrar novos padrões. Um *Data Mining* promove a possibilidade de se transformar dados brutos e sem utilidades de maneira isolada em informação com utilidade para processos que envolvam tomadas de decisão, tornando possível o levantamento hipóteses através de métodos de indução, e os resultados são expressos na forma de regras estatísticas, grafos, árvores de decisão e gráficos. [NAVEGA, 2002]

Motivação e Justificativa

A internet atualmente é a base que constitui maior fonte de informações existente no mundo e a quantidade de dados gerados cresce exponencialmente. Organizações, instituições de ensino, empresas e até mesmo usuários em seu dia a dia, acumulam enorme quantidades de dados, mas apenas acumular dados não é o suficiente para se aproveitar de forma eficiente esses dados em sua totalidade, é necessário que esses dados sejam trabalhados de uma forma com que se tornem conhecimento valioso para que sejam uteis em um processo de tomada de decisão. Coletar, gerenciar e fazer um uso mais eficientes dessas informações exigem técnicas cada vez mais rápidas e inteligentes.

Possuindo quantidades de conteúdo imensurável e possibilidades ainda não exploradas, a web tem se mostrado como um excelente objeto de estudo na área de mineração de dados em massa. Essa quantidade imensa de dados torna inviável a análise humana em um tempo praticável, tornando imprescindível uma ferramenta simples e intuitiva capaz de realizar a mineração dessa enorme de dados e fornecer ao usuário relações importantes a cerca dessas informações para que o mesmo tenha a capacidade de tomar decisões cada vez mais assertivas nas mais diversas áreas.

Objetivos Gerais

O objetivo principal deste trabalho é desenvolver um protótipo para extração e obtenção de conhecimento através de algoritmos de mineração de dados e computação estatística, com estudo de caso à partir de dados postados na web de usuários do microblog *Twitter*. Para se atingir a principal meta do trabalho, os objetivos específicos a seguir deverão ser alcançados:

- Realizar um levantamento bibliográfico dos métodos estatísticos, representações gráficas e algoritmos de extração de conhecimento a serem utilizados;
- Selecionar os termos utilizados e informações providas pela API do *Twitter* para o estudo;
- Adquirir uma coleção de dados para a realização do processamento desejado;
- Implementar o protótipo, utilizando as ferramentas *WEKA* e Linguagem R de forma integrada;
- Testar a aplicação em funcionamento, e avaliar os resultados obtidos

Ao concluir este projeto, espera-se obter uma aplicação que possua um potencial para reconhecer padrões inter-relacionados em grandes bases de dados, e apresentando de forma gráfica e intuitiva os dados devidamente minerados, sendo capaz de permitir ao usuário uma maior eficiência ao analisar dados para a tomada de decisão independente da área de atuação, prevenindo possíveis problemas e capaz de gerar novos conhecimentos através dos métodos implementados durante a execução do projeto da aplicação.

Sendo assim, possível explorar na prática todas as etapas envolvidas durante o desenvolvimento de uma aplicação de *Data Mining* de forma plena e efetiva. Comprovando a importância de sua utilização em fontes massivas de dados (Big Data) tanto para análise preditiva ou para melhorar os processos internos de uma empresa ou organização através da extração e geração de conhecimento.

Fornecendo também uma avaliação das ferramentas visando esclarecer as principais vantagens de se utilizar a Linguagem R e a ferramenta WEKA, para se construir ou projetar uma aplicação utilizando-se de conceitos de extração de conhecimento em bases de dados,

para se efetuar mineração de dados, geração e obtenção de conhecimento em tempo ágil dentro do contexto Big Data, fornecendo relações e representações em forma gráfica e por fim avaliar os resultados alcançados, e as contribuições alcançadas com os estudos, pesquisas, desenvolvimento da aplicação e a análise dos resultados para diversas áreas do conhecimento, verificando se as mesmas foram satisfatórias no caso de estudo.

Organização do Trabalho

Para facilitar a apresentação deste trabalho, ele foi estruturado da seguinte forma:

Os Capítulos 1, 2 constituem toda a parte relacionada ao Referencial Teórico do trabalho. Onde no capítulo um são abordados os conceitos de Big Data e KDD (Extração de Conhecimento) e estabelecido uma relação entre os dois. No capítulo 2 é abordado o conceito de Data Mining, sua estruturação e os métodos para análise das informações resultantes de um processo de mineração.

O Capítulo 3 descreve como foi desenvolvido o projeto de software após as revisões bibliográficas, abordando a metodologia adotada, ambiente de desenvolvimento a interface gráfica com o usuário e os testes realizados.

O Capítulo 4 apresenta os resultados obtidos e as conclusões a respeito da validade e contribuições do trabalho, bem como as possibilidades para trabalhos futuros.

1 Big data e Extração de Conhecimento

Durante este capítulo serão apresentadas as definições dos conceitos de BIG DATA, e Extração de Conhecimento de uma maneira sucinta sobre como são constituídas essas abordagens e estabelecer uma relação entre mesmos.

1.1 Big Data

Para melhor entender o conceito de Big Data, existe uma explicação simples que permite um entendimento e torna o contexto no termo qual está inserido o termo um pouco mais claro. A explicação consiste na seguinte fórmula: O Big Data pode ser conceituado ou representado através de 5 "V"s: volume, variedade, velocidade, veracidade e valor. [TAURION, 2013]

- Volume: Diariamente são gerados quantidades enormes de dados. Sejam em sistemas empresariais ou mesmo usuários através de computadores pessoais ou dispositivos móveis.
- Variedade: Esses dados gerados são provenientes de sistemas estruturados e não estruturados. Os dados de sistemas estruturados são minoria, enquanto os não estruturados fazem parte da imensa maioria desses dados gerados através de envio de e-mails, redes sociais (*Twitter*, *Facebook*, *Blog's*, *Youtube* e outros), documentos eletrônicos, câmeras de vídeo, fotos, mensagens de voz e etc.
- Velocidade: Muitas vezes, para se fazer uso de uma maneira eficaz, é necessário que essas informações sejam analisadas praticamente em tempo real.
- Veracidade: É necessário ter a certeza de que os dados analisados são autênticos e tenham segurança para que as conclusões e tomadas de decisões provenientes das informações geradas através da análise desses dados sejam assertivas.
- Valor: Para se implementar um projeto relacionado a Big Data em uma

> organização (empresa, comércios, área acadêmica, saúde e etc) é absolutamente necessário que a solução implementada traga retorno dos investimentos feitos, do contrário isso resultaria em um grande prejuízo e desperdício de recursos da organização. Um exemplo se aplica na área de seguros, onde uma análise de fraudes pode se tornar mais ágil e imensamente melhor, minimizando riscos, fazendo o uso de análise de dados que não se encontram nas fontes de dados estruturadas que as seguradoras fornecem, como por exemplo dados que estão circulando em diversas mídias sociais. [TAURION, 2013]

1.2 Extração de Conhecimento

O processo de descoberta de conhecimento em bases de dados, segundo Han e Kamber (2006, p. 7) consiste em um método não trivial de extração de informação previamente não conhecida e potencialmente útil dos dados de grandes bases. Com esse processo de descoberta de conhecimento são gerados inúmeros benefícios, como diz Pinheiro (2008, p. 99) "Os maiores benefícios pela execução de processos de mineração de dados é a criação de inteligência de negócios sobre determinado assunto. Também referenciado como KDD - *Knowledge Discovery in Databases* – ou descoberta de conhecimento sob bases de dados."

O processo de extração de conhecimento (KDD) envolve várias etapas, tais como preparação, seleção dos dados, avaliação de padrões, dentre outras. Assim, passando por todas as fases é alcançado o conhecimento, que também poderá visualizado de uma forma objetiva para que o usuário interprete facilmente. De acordo com Han e Kamber (2006, p. 7), cada passo possui funções bem definidas para ser possível chegar ao conhecimento, são elas:

1. Limpeza de dados: realiza a remoção de ruídos, valores nulos e dados inconsistentes.
2. Integração de dados: combina múltiplas fontes de dados, quando necessário, em uma única fonte coerente de dados.
3. Seleção de dados: seleciona os dados da base que serão relevantes para realizar a análise.
4. Transformação de dados: transforma os dados em grupos apropriados para a etapa de mineração realizando, por exemplo, processos de sumarização ou agregação.

5. Data Mining: aplica métodos inteligentes para extrair padrões dos dados de forma automática.
6. Avaliação de padrões: baseia-se em medidas para visualizar os padrões que representam conhecimento interessante obtido na etapa de Data Mining.
7. Apresentação do conhecimento: utiliza técnicas de representação e visualização para apresentar ao usuário o conhecimento extraído nos dados existentes.

A sequência dos passos necessários para completar o processo de descoberta de conhecimento em base de dados é demonstrada na figura a seguir:

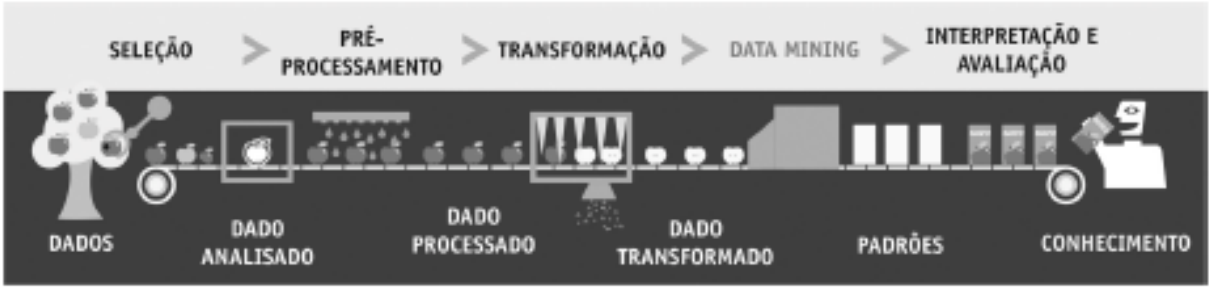

Figura 1. Processo de KDD(Fayyad, apud, Gonçalves, 2001)

1.3 Relação entre Big Data e Extração de Conhecimento

Ao trazer os conceitos de extração e descoberta de conhecimento, para dentro do contexto de Big Data, podemos abrir uma nova gama de possibilidades para se criar ferramentas que integrem os dois conceitos para se atender uma necessidade específica ou até mesmo identificar necessidades anteriormente não identificadas, de forma a auxiliar tomadas de decisão, podendo resultar em melhoria de processos em organizações, mudanças nas estratégias coorporativas e apresentar resultados de uma maneira mais rápida, eficiente e assertiva.

Existem maneiras de se relacionar esses conceitos de forma prática através de conceitos já estudados, como a mineração de dados. Existem diversas soluções disponíveis de forma gratuita, como por exemplo, utilizar-se de ferramentas fornecidas pela tecnologia da informação para se construir uma solução de Data Mining é uma forma de se explorar de forma prática a relação entre os conceitos abordados durante o projeto.

2 Mineração de Dados

Durante este capítulo serão apresentadas as definições dos conceitos de Data Mining (Mineração de Dados) e de uma maneira sucinta e definir teoricamente as etapas da construção de uma ferramenta para realizar o trabalho de mineração de dados.

2.1 Data Mining

Data Mining (Mineração de Dados) trata-se da etapa mais importante na extração de conhecimento. É nessa etapa que os dados são efetivamente transformados em conhecimento através de algoritmos que exploram esses dados em buscas de associações entre variáveis, e estabelecendo relações entra as mesmas, revelando padrões que podem ser importantes na interpretação desses dados.

- Classificação – arruma os dados em grupos predefinidos.
- Clustering – arruma os dados em grupos não predefinidos encontrando padrões de semelhança entre os dados.
- Regressão – tenta encontrar uma função que modele os dados com o menor índice de erros possíveis.
- Aprendizado de regras de associação – procura por relacionamentos entre as variáveis.

As ferramentas utilizadas para se fazer um processo de mineração de dados utilizam-se de uma série de princípios que levam a conclusões através de processos de indução, diferente das técnicas comuns de banco de dados que oferecem uma conclusão dedutiva, que são inadequadas para se encontrar novos padrões. [NAVEGA, 2002]

Um Data Mining promove a possibilidade de se transformar dados brutos e sem utilidades de maneira isolada em informação com utilidade para tomadas de decisão, é possível se levantar hipóteses através da indução, e os resultados são expressos na forma de regras estatísticas, grafos, arvores de decisão. [NAVEGA, 2002]

Recursos de mineração de dados podem ser utilizados em uma infinidade de áreas, promovendo resultados satisfatórios em diversos casos [Olson, 2008 e Witten, 2005], como

por exemplo:

- Prevenção: Detectar fraudes ou crimes, possibilitar médicos a darem diagnósticos mais precisos em um tempo menor, verificar probabilidade de prejuízo em operações, evitar acidentes de transito ou situações que coloquem uma pessoa em risco antes mesmos que elas aconteçam através de uma análise preditiva.
- Aumento da rentabilidade de negócios: identificar inconsistências e oportunidades de melhoria em processos, redução de despesas selecionando os melhores fornecedores e serviços terceirizados de uma empresa.
- Fidelização de clientes: possibilita identificar perfis de consumidores de um determinado produto, relações entre a compra de vários produtos e melhorar gerenciamento de relacionamento com o cliente.
- Cientifica: Colaboração em pesquisas cientificas (como na área de inteligência artificial, ou reconhecimento de padrões em códigos genéticos)

2.2 Etapas de construção de um Data Mining

De acordo com Han e Kamber (2006, p. 8) a arquitetura básica de um sistema elaborado para se efetuar a mineração de dados é apresentado da seguinte forma:

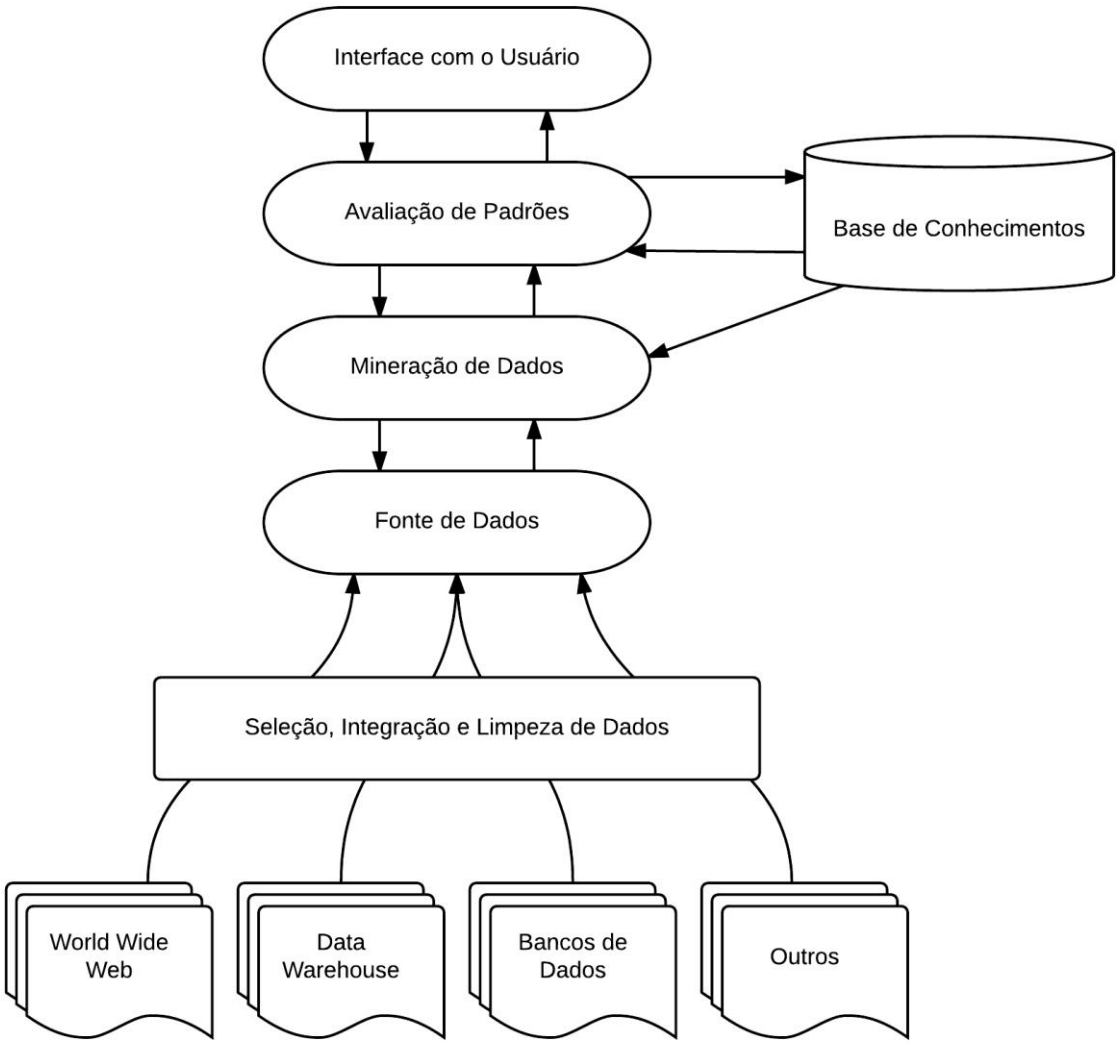

Figura 2. Arquitetura básica de um sistema de Data Mining (HAN, KAMBER 2006)

No processo de avaliação de padrões, uma ferramenta de Data Mining utiliza-se de algoritmos que nada mais são que mecanismos que criam os modelos para a mineração. Para se criar este modelo, o algoritmo realiza uma análise de um conjunto de dados e identifica padrões inter-relacionados entre variáveis e também encontra tendências nos dados. Dentre os vários modelos de algoritmos usados nas ferramentas de Data Mining destacam-se:

- Algoritmos para classificação dos dados;
- Algoritmos para regressão;
- Algoritmos para segmentação;
- Algoritmos para associação de variáveis;
- Algoritmos para análise de sequências;

2.3 Métodos de Análise

Para se apresentar os dados, na hora de promover facilidade de interpretação do conhecimento extraído das bases de dados, é necessário que a apresentação seja facilitada, de forma com que o usuário possa analisar de forma intuitiva e tomar decisões assertivas.

Para isso, uma atenção especial é voltada para a parte de interface com o usuário no qual deve apresentar a visualização dos resultados obtidos, pois em grande parte, a informação extraída não é muito clara. A melhor forma de estabelecer esse relacionamento dos conhecimentos obtidos com o usuário é através de gráficos, tabelas e outras formas visuais que devem ser exploradas com o objetivo de facilitar a interpretação de saídas da ferramenta de mineração de dados.

Os mais utilizados, por gerar uma maior facilidade de interpretação para pessoas que não possuem um grande conhecimento em estatística são os gráficos em barra/colunas, gráficos de setores, árvores de decisões e tabelas por permitir que com um olhar genérico, possam se tomar diversas linhas de raciocínio e entender os resultados mesmo não possuindo grandes conhecimentos em estatísticas.

Por fim, a etapa de análise dos dados deve verificar se a informação que gerou o conhecimento é importante para o assunto estudado, ou seja, deve ser validada, e deve ter relevância para o objeto de estudo, através de métricas que devem ser quantitativas e análises qualitativas que deverão ser interpretadas por um especialista.

3 MINERAÇÃO DE DADOS PARA OBTENÇÃO DE CONHECIMENTO EM BIG DATA

Dada importância dos estudos de Mineração de Dados, Extração de Conhecimentos e a contextualização do *Big Data*, nota-se a necessidade de uma forma de auxiliar um usuário ou gestor em um processo de tomada de decisão dado a impossibilidade de análise de uma quantidade imensurável de dados. Onde o principal objetivo da mineração de dados, é possibilitar de forma automática de descoberta de conhecimento que se encontra oculta em grandes quantidades de dados, permitindo uma agilidade nas tomadas de decisões.

Diante deste contexto, é interessante que se desenvolva o protótipo de uma ferramenta que forneça uma melhor visualização das informações coletadas durante todo o processo de mineração e extração de conhecimento em bases de dados de grande porte.

3.1 Metodologia

A fim de se atingir os objetivos gerais e específicos deste projeto, em primeiro lugar foi efetuado o levantamento bibliográfico de estudos contemplando os assuntos alvos de abordagem deste trabalho, nos quais serviram de base de fundamentação teórica para a aquisição de elementos que definiram este projeto.

Tomando como base as literaturas, foi traçado um quadro teórico para sustentar o desenvolvimento de toda a pesquisa, consolidando-o e alinhando-o com os objetivos do projeto.

Posteriormente à fundamentação teórica, foi feito o levantamento das tecnologias e ferramentas disponíveis de forma gratuita para a implementação do projeto de software, e escolhidas através de critérios como fácil acesso à documentação e desempenho em relação à outras ferramentas disponíveis, suportadas em sistemas operacionais Linux e Windows, e possibilidade de integração das ferramentas com a linguagem de programação Java, as quais duas ferramentas foram escolhidas como alvo de estudos: WEKA e Linguagem R.

Em seguida, foram definidas as fontes de dados que seriam alvo de estudo para o desenvolvimento da aplicação de mineração de dados, nas quais deveriam estar à disposição de forma gratuita. Para o projeto, foi definida a utilização do microblog Twitter, por oferecer uma API com boa documentação.

Durante as pesquisas chegou-se à conclusão de que devido à facilidade de documentação e fácil integração a Linguagem *JAVA* seria utilizada.

3.2 WEKA

Trata-se de software livre que consistem em uma coleção de algoritmos de aprendizado de máquinas e para tarefas de mineração de dados desenvolvidos na linguagem de programação Java. Os algoritmos podem ser aplicados diretamente à um conjunto de dados ou chamado a partir de um código em um programa escrito em Java. *WEKA* contém ferramentas para pré-processamento de dados, classificação, regressão, clustering, regras de associação e visualização. Também é adequado para o desenvolvimento de novos sistemas para aprendizado de máquina. [Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, 2009]

Weka possui uma interface que possibilita ao usuário utilizar e executar rotinas de *Data Mining*, onde podem ser feitos: pré-processamento, classificação, *clustering*, associação, seleção de atributos e apresentação de gráficos:

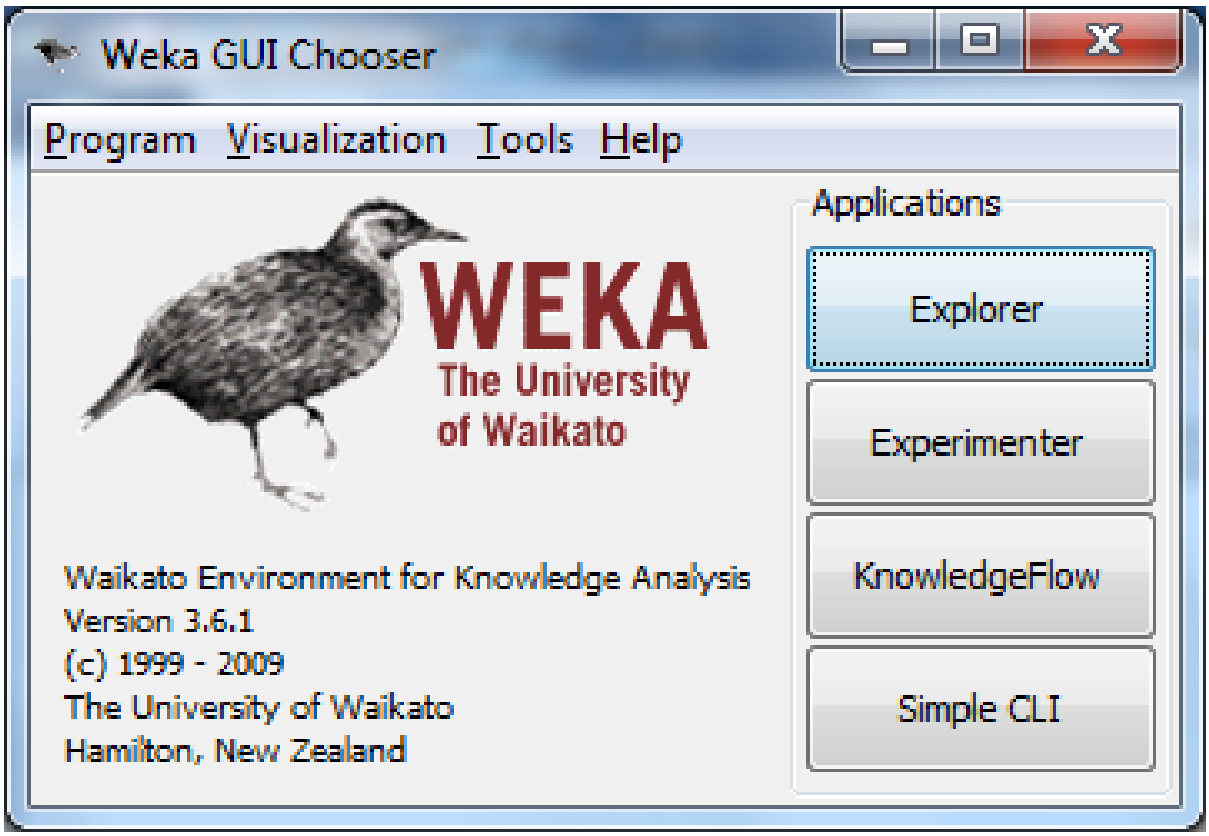

Figura 3. Tela Inicial da Interface com o usuário – WEKA

WEKA também pode ser utilizada através de linha de comando, ou também pode ser utilizada através de seu próprio código fonte, usando a linguagem de programação Java, e pode ser executado nos sistemas operacionais Linux, Windows e Mac.

Para o desenvolvimento deste projeto, foi usada a interface com o usuário para se testar a ferramenta, e foi definida a utilização do seu código fonte para o desenvolvimento de uma aplicação escrita em Java, para utilizar-se dos recursos da linguagem para mineração, pré-processamento, classificação, regras de associação e visualização.

Durante o desenvolvimento, foram definidos os algoritmos e funções que seriam utilizados alguns algoritmos e funções oferecidos. Em WEKA, foi escolhido o algoritmo J48, que é um algoritmo para classificação de dados. Segue um exemplo de gráfico de árvore de decisões resultante do algoritmo J48:

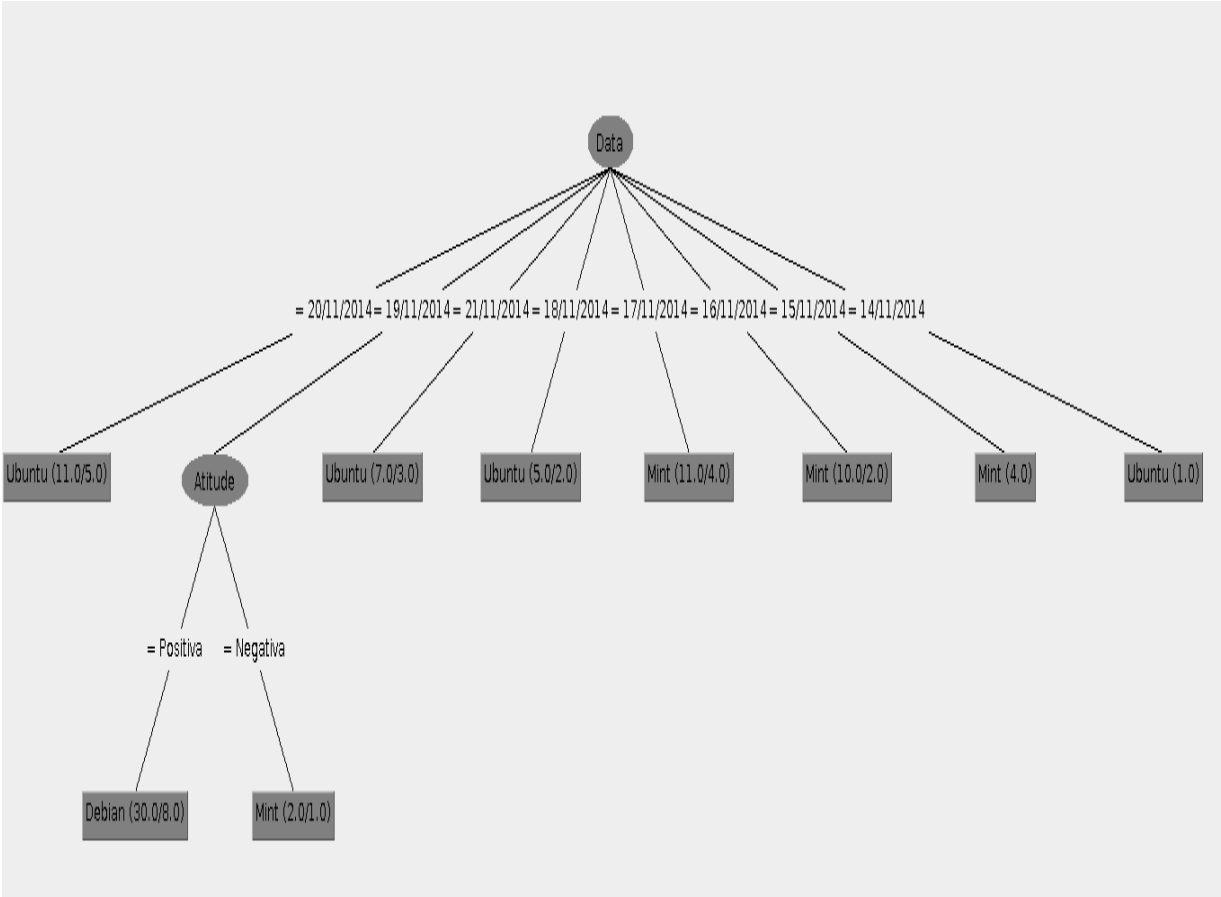

Figura 4. Exemplo de Gráfico de Árvore de Decisões J48.

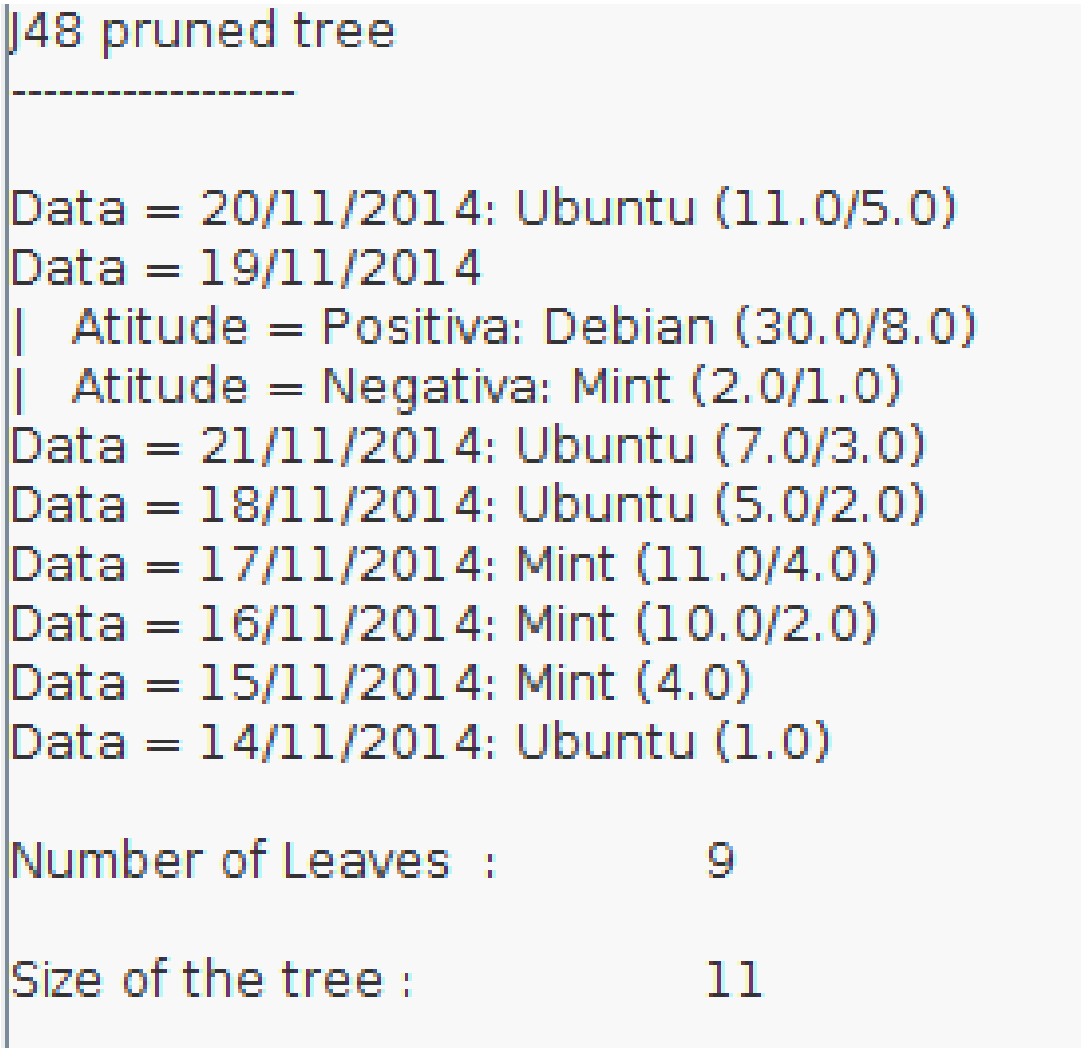

Figura 5. Exemplo textual de Árvore de Decisões J48.

O código desenvolvido para a implementação no protótipo foi o seguinte:

```
360     public String mostraArvoreJ48() throws Exception{
361
362         // Cria o classificador
363         J48 cls = new J48();
364         Instances data = new Instances(new BufferedReader(new FileReader("archives/dados.arff")));
365
366         // Deleta os atributos indesejados pelo indice
367         data.deleteAttributeAt(0);
368         data.deleteAttributeAt(2);
369         data.deleteAttributeAt(7);
370         data.deleteAttributeAt(6);
371         data.deleteAttributeAt(5);
372         data.deleteAttributeAt(3);
373
374         // Indice utilizado na classificação
375         data.setClassIndex(0);
376
377         // Cria a classificação dos dados
378         cls.buildClassifier(data);
379
380         // Mostra em uma janela, uma árvore de decisão resultante da classificação
381         final javax.swing.JFrame jf = new javax.swing.JFrame("TwitGoldMiner - Árvore de Decisão : J48");
382             jf.setSize(800,600);
383             jf.setLocationRelativeTo(null);
384             jf.getContentPane().setLayout(new BorderLayout());
385             TreeVisualizer tv = new TreeVisualizer(null, cls.graph(), new PlaceNode2());
386             jf.getContentPane().add(tv, BorderLayout.CENTER);
387             jf.addWindowListener(new java.awt.event.WindowAdapter() {
388               public void windowClosing(java.awt.event.WindowEvent e) {
389                 jf.dispose();
390               }
391             });
392
393             jf.setVisible(true);
394             tv.fitToScreen();
395
396         return(cls.toString());
397     }
398 }
```

Figura 6. Código implementado - Algoritmo J48.

3.3 Linguagem R

Algumas ferramentas para se fazer análises estatísticas de dados computacionais e representação em forma gráfica disponíveis como Software Livre sobre os termos da GNU com licença GPL (General Public License), dentre elas a Linguagem R se destaca demonstrado uma certa superioridade em vários fatores, com relação à outras ferramentas para análise estatísticas disponíveis gratuitamente. Como demonstra a imagem a seguir:

	DESC	FREQ	PROB	ANOVA1	ANOVA+	EXPER	SLR	MLR	LOG	LOGIT	PROBIT	GLM	ANCOVA	NONPAR	LOGLIN	TIME	SURV	PCA	FACT	CCA	CA	DISCR	CLUST
DATAPLOT	✔	✔	✔	✔	✔	✔	✔	✔						✔		✔		✔		✔		✔	
EASYREG	✔	✔	✔				✔	✔		✔	✔	✔		✔		✔	✔						
GRETL	✔	✔	✔				✔	✔	✔	✔	✔	✔		✔		✔		✔					
INSTAT+	✔	✔	✔	✔	✔		✔	✔						✔	✔	✔	✔						
MACANOVA	✔	✔	✔	✔	✔		✔	✔	✔	✔		✔	✔			✔		✔	✔				✔
MICROSIRIS	✔	✔	✔	✔	✔		✔	✔		✔					✔		✔	✔	✔				✔
R	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
TANAGRA	✔			✔			✔	✔	✔					✔				✔	✔		✔	✔	✔
VISTA	✔	✔		✔	✔		✔	✔							✔			✔			✔		✔
WINIDAMS	✔	✔		✔			✔	✔								✔		✔	✔			✔	✔

Figura 7. Ferramenta Estatística Gratuita.(FREE STATISTICS, 2014)

LEGENDA:

DESC: Descriptive Statistics
FREQ: Distribution Frequencies
PROB: Probability Distributions
ANOVA1: One Way Analysis of Variance
ANOVA+: Two or Three Way Analysis of Variance
EXPER: Experimental Design
SLR: Simple Linear Regression
MLR: Multiple Linear Regression
LOG: Logistic Regression
LOGIT: Logit Model
PROBIT: Probit Model
GLM: Generalized Linear Models
ANCOVA: Analysis of Covariance
NONPAR: Non Parametric Test
LOGLIN: Log-Linear Analysis
TIME: Time Series
SURV: Survival Analysis
PCA: Principal Component Analysis
FACT: Factorial Analysis
CCA: Canonical Correlation Analysis
CA: Correspondence Analysis
DISCR: Discriminant Analysis
CLUST: Cluster Analysis

Figura 8. Legenda – Estatística Gratuita.(FREE STATISTICS, 2014)

Trata-se de um ambiente ou linguagem de programação para computação estatística e gráficos. O R fornece uma variedade recursos para análises estatísticas e apresentação em forma gráfica, sendo assim uma ferramenta e um conjunto integrado de facilidades de software para a manipulação de dados, cálculo e exibição gráfica [R-PROJECT, 2014]. E inclui:

- Uma manipulação de dados eficaz e instalação de armazenamento.
- Um conjunto de operadores para cálculos em matrizes, em especial as matrizes.
- Uma grande coleção, coerente e integrada de ferramentas intermediárias para análise de dados, instalações gráficas para análise de dados e visualização, quer na tela ou em cópia impressa.
- Uma linguagem de programação simples e eficaz bem desenvolvida que inclui condicionais, laços, funções recursivas definidas pelo usuário e recursos de entrada e saída. [R-PROJECT, 2014]

Uma das vantagens de se utilizar a Linguagem R para se construir gráficos é a velocidade com que ela associa as varáveis informadas e gera o gráfico de forma extremamente rápida. Como no exemplo a seguir: associando os valores de “Busca” e “Atitude”:

```
1  #----- Lê os dados do arquivo csv
2  dados <- read.table("archives/dados.csv", head=T, sep=",")
3  attach(dados)
4
5  #----- Cria as tabelas com as informações para utilizar nos gráficos
6  busca.atitude <- table(Busca,Atitude)
7  atitude.busca <- table(Atitude,Busca)
```

Figura 9. Código de Associação de Variáveis em Linguagem R.

Para o desenvolvimento do protótipo as funções escolhidas da Linguagem R para serem utilizadas no software, foi escolhida a função “PIE” (torta) que realiza associações em uma única variável que resulta em um Gráfico por Setores. Exemplo:

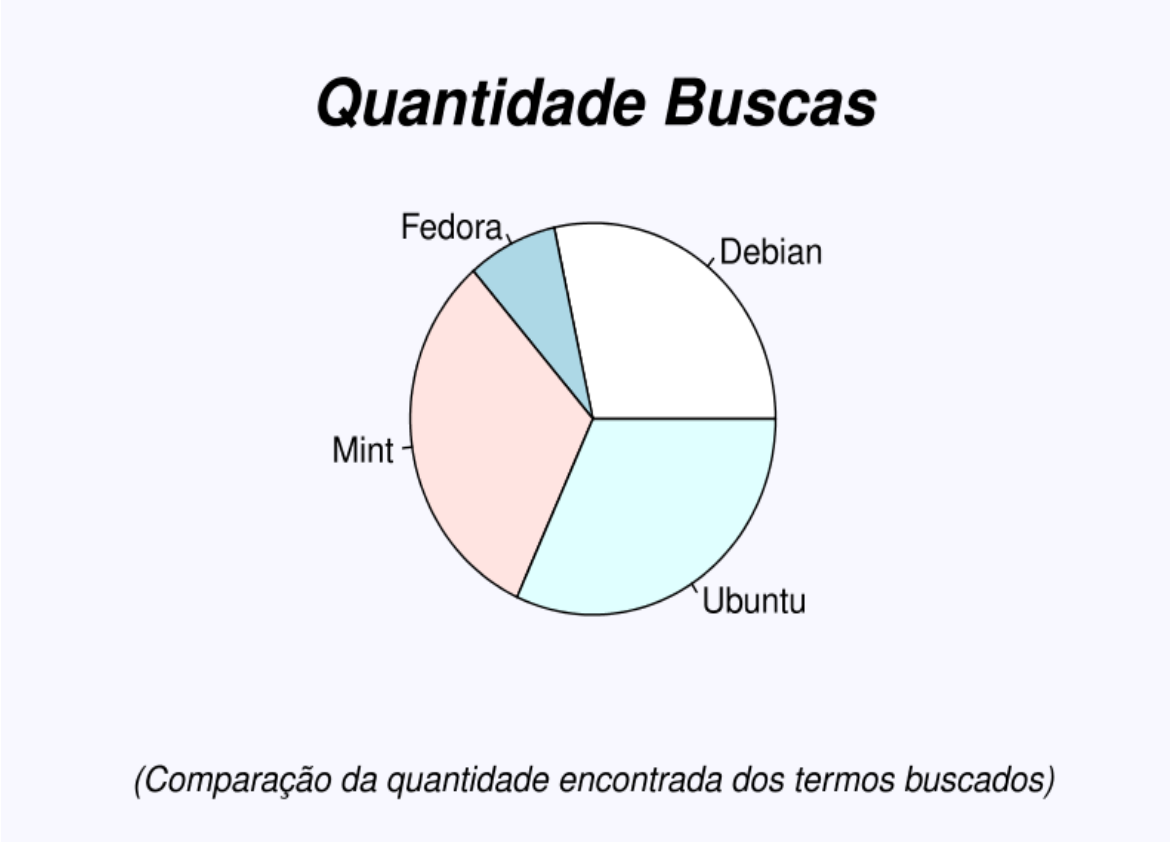

Figura 10. Exemplo de Gráfico de Setores.

Exemplo de código em Linguagem R que resulta no gráfico de setores acima:

```
#----- Cria um pdf para armazenar os gráficos criados
pdf("archives/relatorio.pdf", width=6, height=4)

#----- Gráfico em pizza - Quantidade Buscas
par(bg="GhostWhite")
par(mar=c(5,2,5,2))
par(cex=1)
pie(table(Busca), radius = 1)
title(main = "Quantidade Buscas", xlab = "(Comparação da quantidade encontrada dos termos buscados)",cex.main = 1.8, cex.lab = 1.0, font.main = 4, font.lab = 3)

#----- Salvo os gráficos no arquivo PDF criado
dev.off()
```

Figura 11. Código para geração de Gráfico de Setores em Linguagem R.

Também foi utilizada no projeto do protótipo foi a função “BARPLOT” que realiza associações em uma ou mais variáveis que resulta em um Gráfico de Barras. Exemplo:

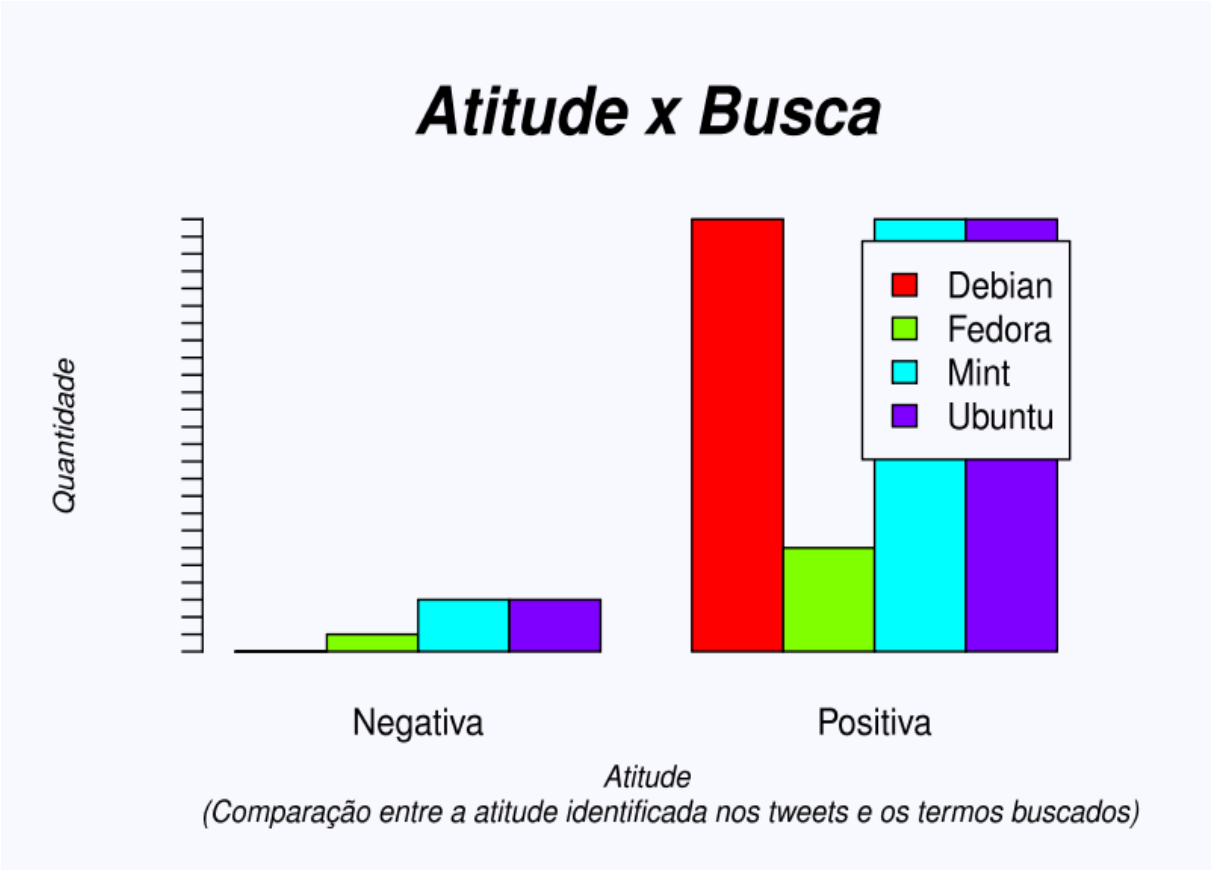

Figura 12. Exemplo Gráfico de Barras.

Exemplo de código em Linguagem R que resulta no gráfico em barras acima:

```
#----- Cria um pdf para armazenar os gráficos criados
pdf("archives/relatorio.pdf", width=6, height=4)

#----- Gráfico em barras - Atitude x Busca
par(bg="GhostWhite")
par(mar=c(5,5,5,3))
par(cex=1)
barplot(busca.atitude,axes = F, beside = T, legend = T,col = rainbow(???))
title(main = "Atitude x Busca", xlab = "Atitude
(Comparação entre a atitude identificada nos tweets e os termos buscados)", ylab = "Quantidade", cex.main = 1.8, cex.lab = 0.8, font.main = 4, font.lab = 3)
axis(2, at=x, label = F)

#----- Salvo os gráficos no arquivo PDF criado
dev.off()
```

Figura 13. Código para geração de Gráfico de Barras em Linguagem R.

3.4 Ambiente de Desenvolvimento

A concepção do software foi realizada utilizando o sistema operacional Linux Mint e a Eclipse – IDE, todo o projeto foi escrito em linguagem de programação JAVA em conjunto com a Linguagem R e a WEKA, também fazendo uso de algumas bibliotecas externas, e para persistir as informações pesquisadas pelo usuário, foi utilizado o banco de dados Firebird 2.5.

Bibliotecas Externas utilizadas durante a implementação do protótipo de software para mineração dados:

- TWITTER 4J. (http://twitter4j.org)
- CSVReader. (http://www.csvreader.com/)
- PDFRenderer. (https://java.net/projects/pdf-renderer)

3.5 Interface da Aplicação

Ao iniciar a aplicação, o usuário irá visualizar a janela de Autenticação, que dará acesso às demais funcionalidades do protótipo, para isso informando o usuário e senha previamente inseridos no banco de dados.

Figura 14. Tela Autenticação.

Após isso, o usuário terá visibilidade para todas as funções do protótipo a partir da tela principal, na qual se encontra a aba Mineração, selecionada na opção de “Pesquisa”.

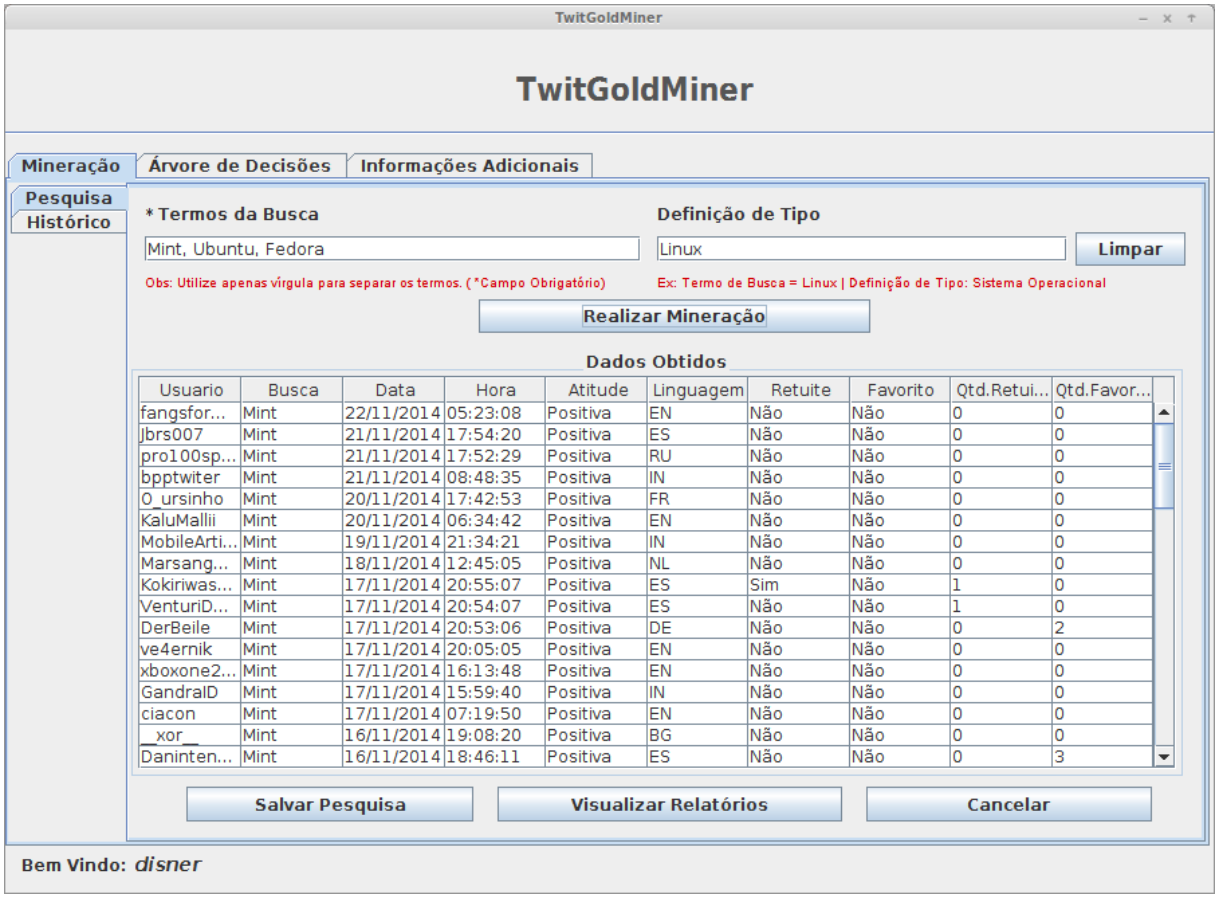

Figura 15. Tela Mineração – Pesquisa.

Para realizar uma mineração o usuário deverá fornecer os termos da busca separados por vírgula, e se desejar definir um "Tipo" para a busca. Após isso clicar no botão "Realizar Mineração". Ao realizar a mineração o usuário irá visualizar as informações extraídas do *Twitter* em uma tabela e serão habilitadas as opções "Salvar Pesquisa", "Visualizar Relatórios" e a opção "Visualizar Árvore" na aba "Árvore de Decisões".

Ao selecionar a opção "Salvar Pesquisa" a ferramenta irá gravar a pesquisa no banco de dados que poderá ser consultada novamente através da opção "Histórico" e reiniciar a tela, possibilitando que seja efetuada uma nova pesquisa.

Ao selecionar a opção "Visualizar Relatórios", a aplicação irá mostrar ao usuário a tela de visualização de relatórios.

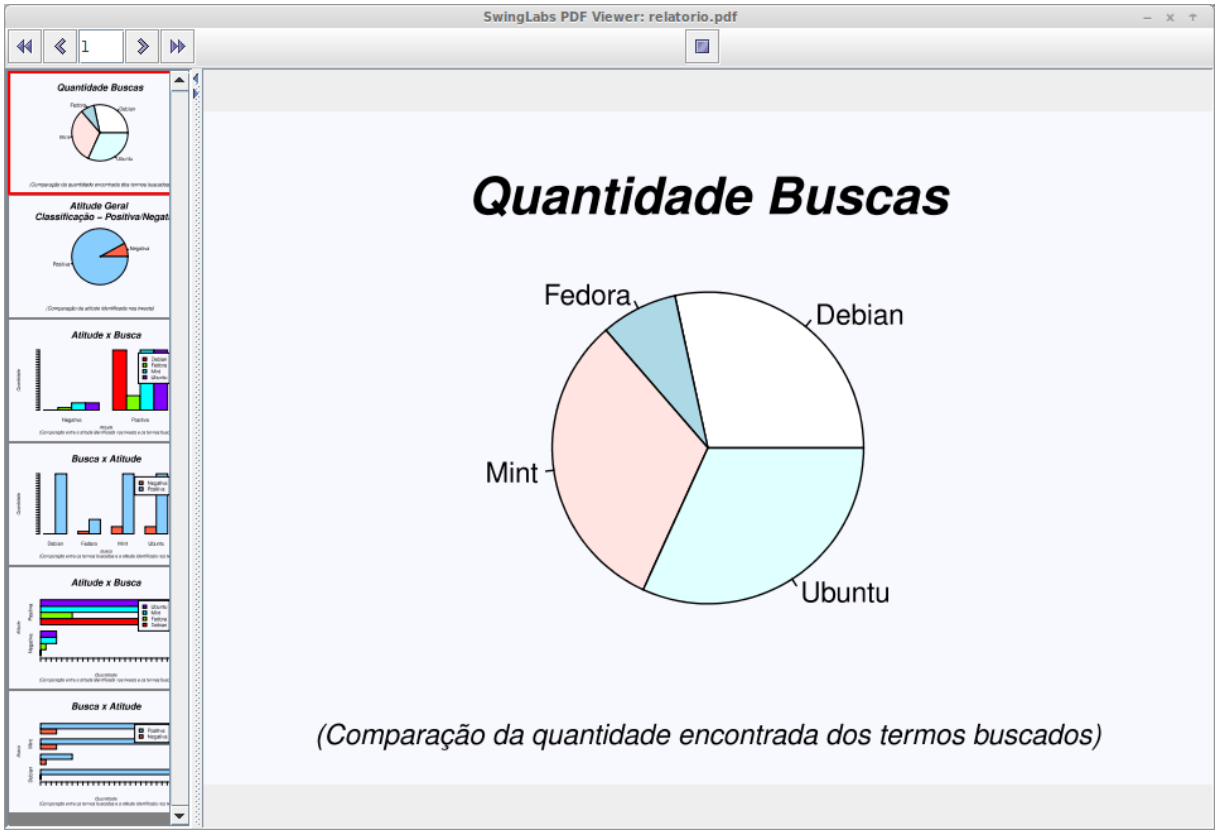

Figura 16. Tela Visualizar Relatórios.

O usuário pode também selecionar a aba “Árvore de decisões” e selecionar a opção “Visualizar Árvore”, que irá mostrar em uma nova janela um gráfico da Árvore de Decisões gerada.

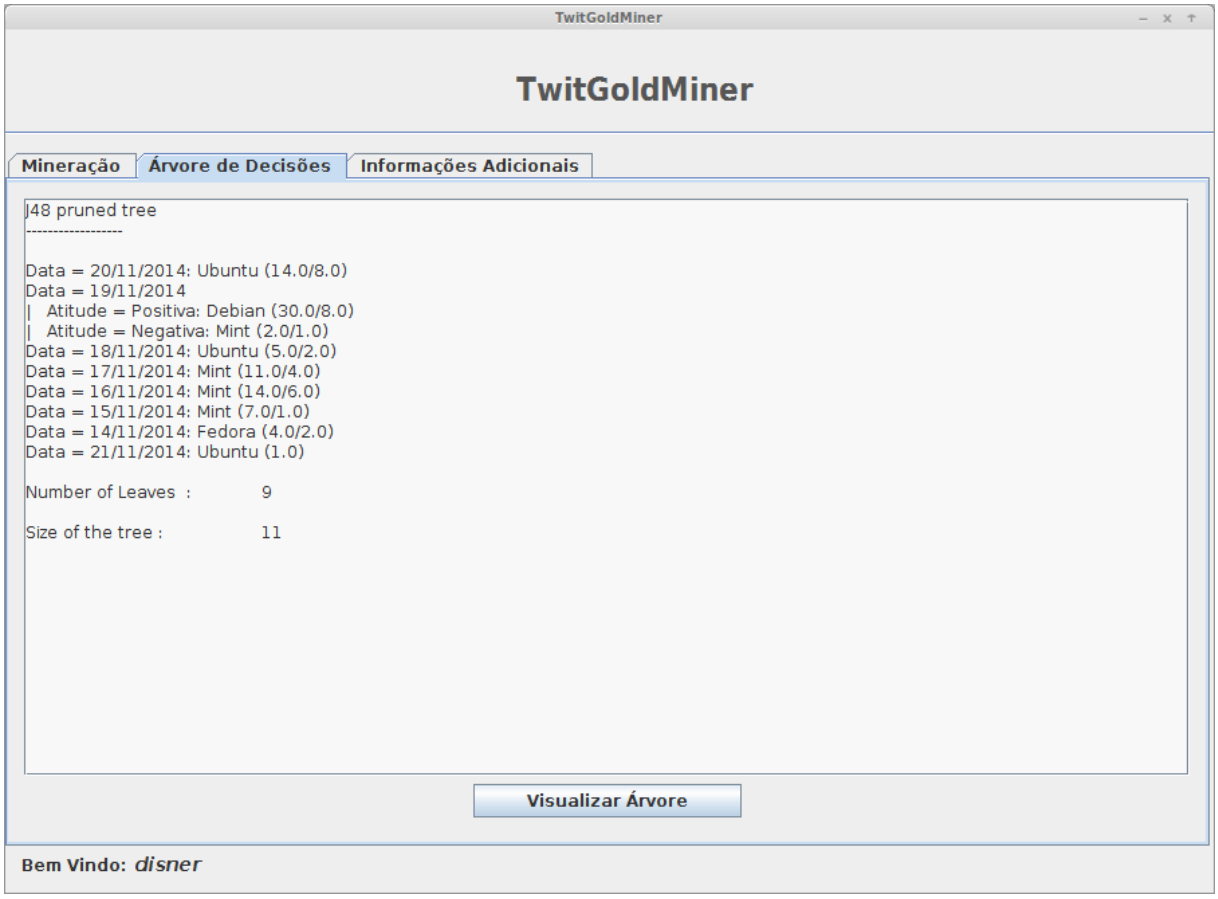

Figura 17. Tela Árvore de Decisões.

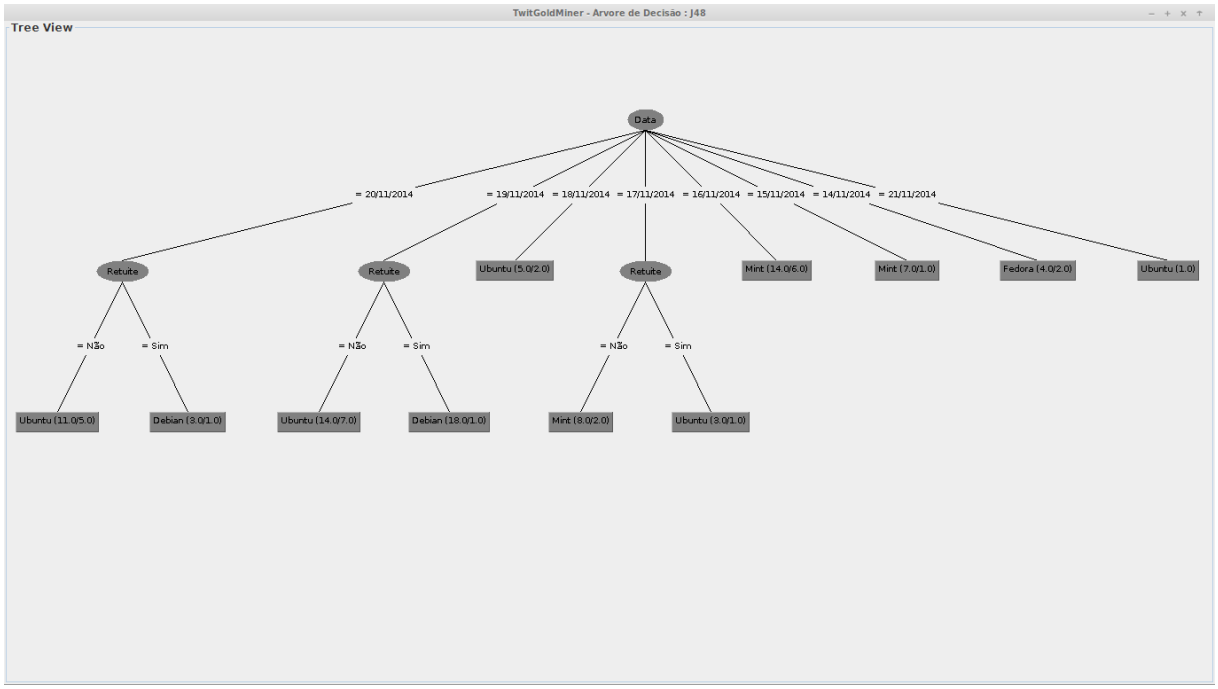

Figura 18. Tela Visualizar Gráfico - Árvore de Decisões

Caso o usuário deseje, ele pode visualizar outras pesquisas feitas através da opção “Histórico”

na aba “Mineração”. Onde poderá escolher em uma lista, todas as opções de pesquisas efetuadas e salvas, selecionando a pesquisa e clicando no botão “Visualizar Pesquisa”, tendo as opções de visualizar os relatórios e árvore de decisões habilitadas.

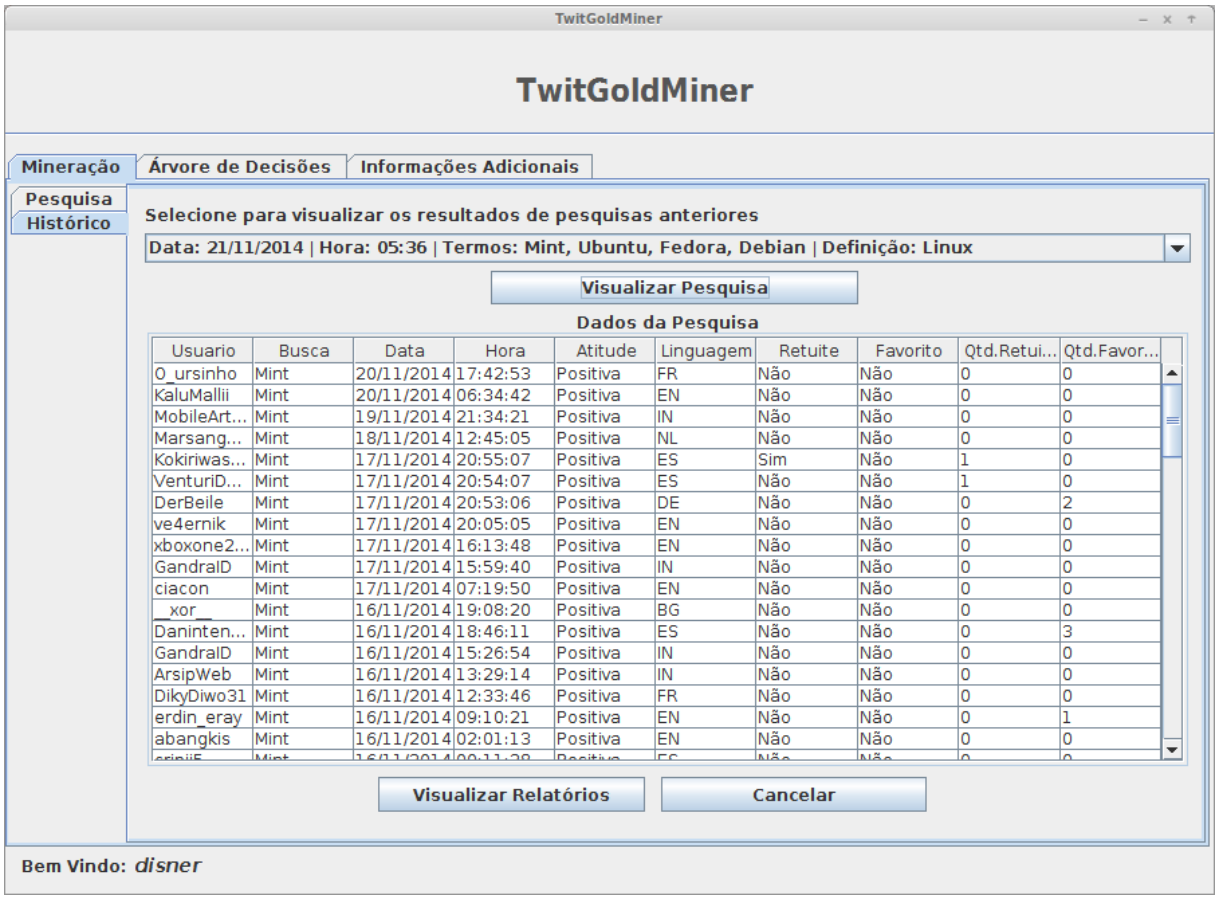

Figura 19. Tela Mineração - Histórico

3.6 Testes Efetuados

Foram efetuados testes, comparando a popularidade de algumas distribuições Linux através dos gráficos comparando a quantidade de opiniões negativas e positivas de cada uma e a quantidade total de opiniões positivas e negativas. Como demonstram os gráficos a seguir:

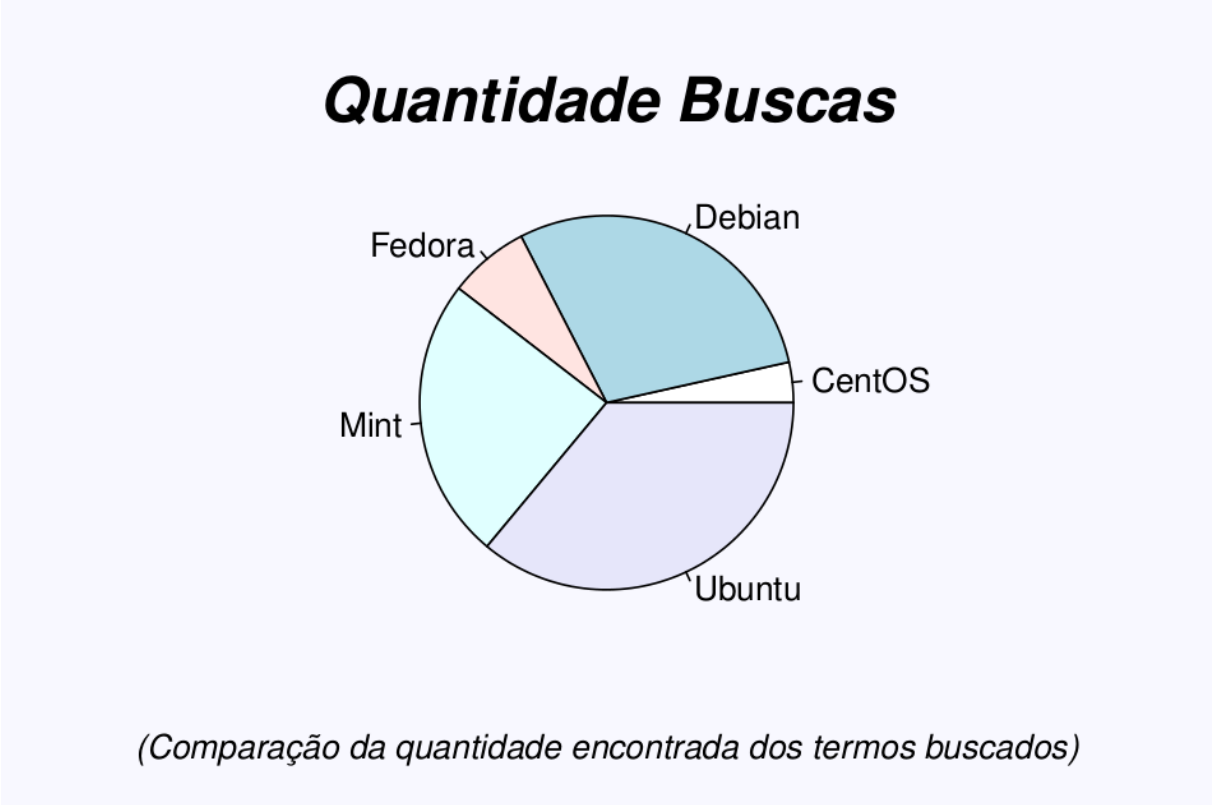

Figura 20. Testes - Distribuições Linux (Quantidade de Buscas)

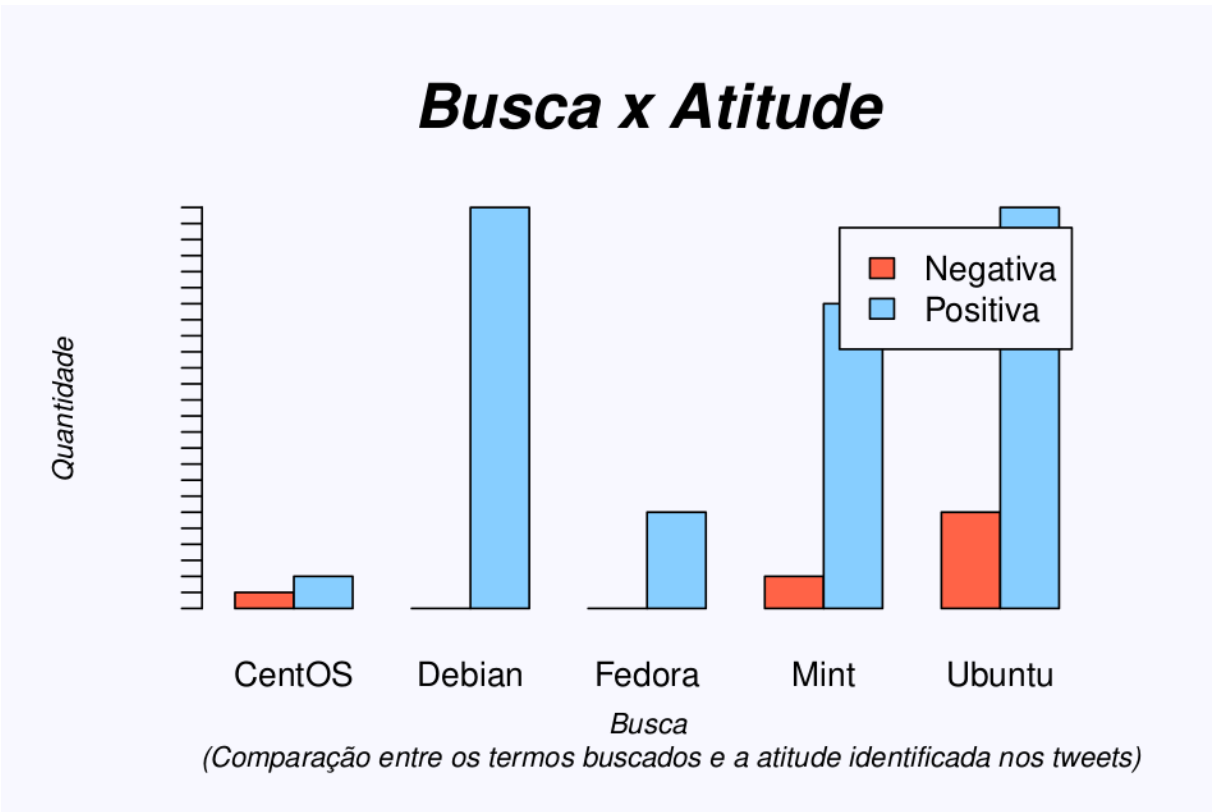

Figura 21. Testes – Distribuições Linux (Busca x Atitude)

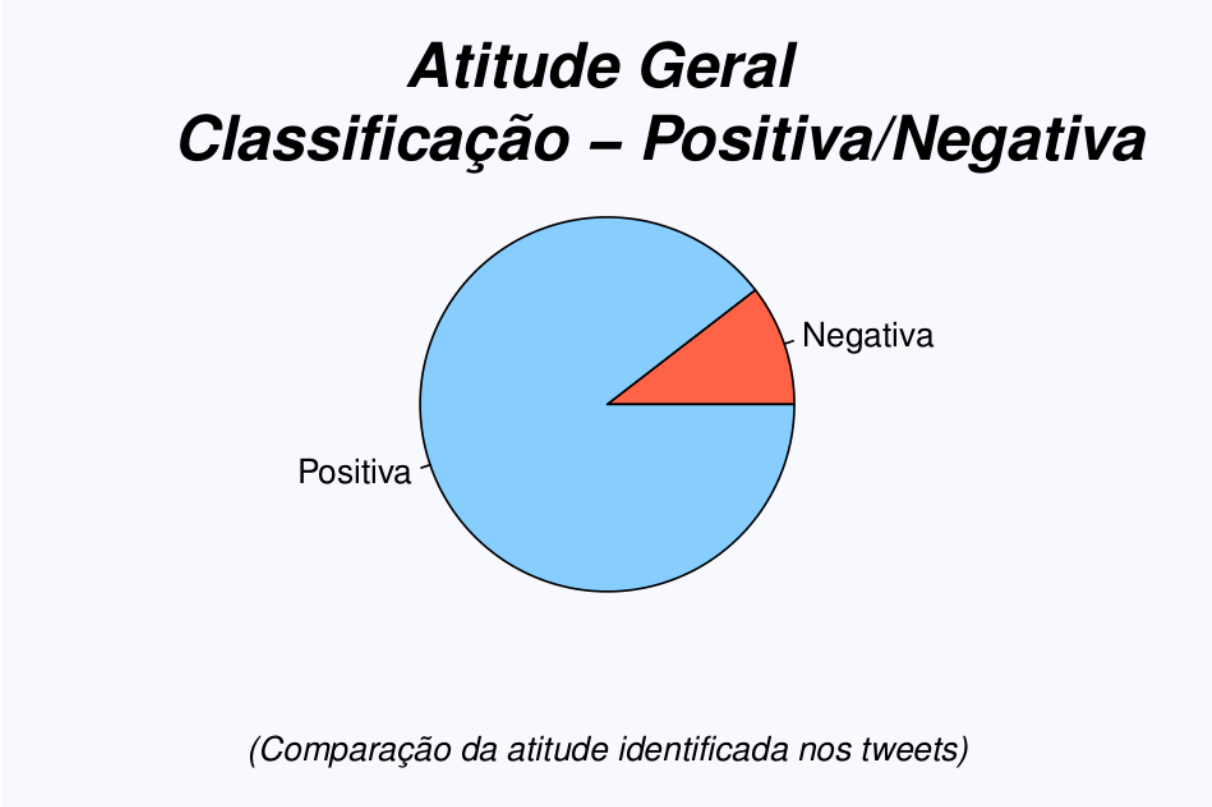

Figura 22. Testes – Distribuições Linux (Atitude Geral)

As árvores de decisões geradas pelo protótipo, também permitiram avaliar as opiniões em uma perspectiva diferente, associando os termos buscados e as atitudes dos usuários em relação aos termos buscados, com a data em que foi feita a postagem do conteúdo no *Twitter*.

```
J48 pruned tree
------------------

Data = 24/11/2014: Ubuntu (10.0/3.0)
Data = 23/11/2014: Ubuntu (12.0/3.0)
Data = 22/11/2014: Ubuntu (7.0/2.0)
Data = 21/11/2014: Ubuntu (8.0/4.0)
Data = 20/11/2014: Ubuntu (12.0/8.0)
Data = 19/11/2014
|   Atitude = Positiva: Debian (21.0/2.0)
|   Atitude = Negativa: Ubuntu (2.0/1.0)
Data = 18/11/2014: Mint (4.0/2.0)
Data = 17/11/2014: Mint (7.0)
Data = 16/11/2014: Mint (3.0)

Number of Leaves  :         10

Size of the tree :          12
```

Figura 23. Testes - Distribuições Linux (Árvore Decisões Textual)

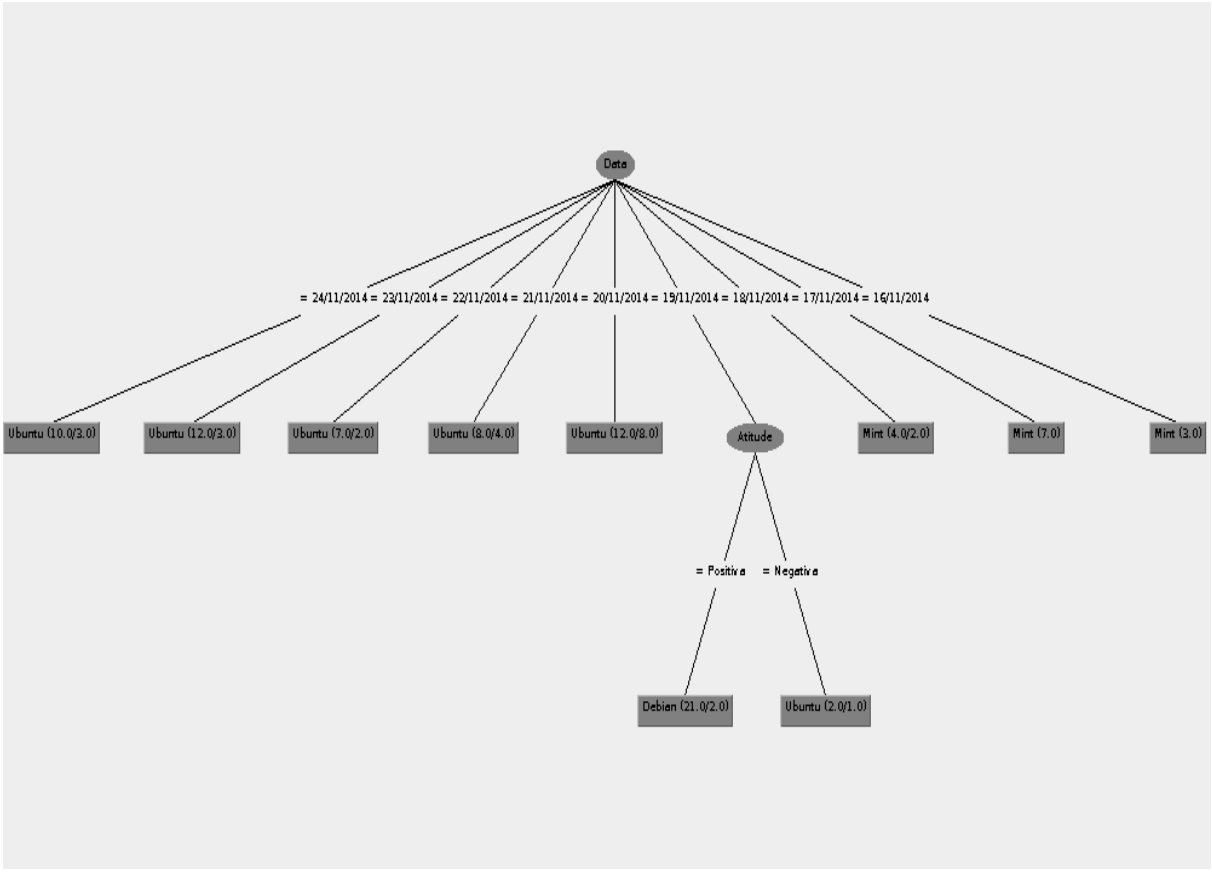

Figura 24. Testes - Distribuições Linux (Árvore Decisões Gráfico)

Os mesmos testes também foram feitos com algumas marcas conhecidas de fabricantes de diversos produtos como notebooks, celulares, carros, roupas e artigos esportivos. Chegando ao mesmo tipo de conclusão, acerca da popularidade e opiniões sobre os termos buscados postados pelos usuários do *Twitter*.

4 Resultados e Contribuições

Após a implementação, ficou claro, que ao se realizar um estudo dos temas apresentados e através da metodologia aplicada, pode-se concluir que é possível desenvolver uma aplicação para encontrar conhecimento útil nas mais diversas fontes de dados, que ofereça um bom desempenho e entregue ao usuário final, uma interface amigável e intuitiva, sem que haja uma grande alocação de recursos.

Dado a simplicidade do protótipo desenvolvido no projeto, é possível notar que para se aplicar esse s conhecimentos em grandes projetos, voltados para diversas áreas como pesquisa científica e empresarial, torna-se extremamente útil, e aplicável, pois é possível utilizar-se apenas de recursos oferecidos de forma gratuita, ficando a cargo apenas os custos de despesas para o desenvolvimento e implantação de uma ferramenta de Mineração de Dados em bases de dados complexas e com grande volume de dados.

CONCLUSÃO

Ao realizar o estudo sobre os temas envolvidos no trabalho, após a revisão bibliográfica observou-se que os projetos relacionados à Mineração de Dados e *Big Data* são campos promissores, principalmente se associados à *Web*, que fornece uma infinidade de dados que se trabalhados, selecionados e analisados de maneira inteligente e automatizada, possuem um enorme potencial de se transformar em conhecimento útil para organizações ou pessoas.

O grande mérito do trabalho é a conclusão de que através do protótipo de *Data Mining* implementado, foi possível evidenciar a possibilidade de se desenvolver uma ferramenta para extrair conhecimento no contexto *Big Data*. Utilizando para desenvolver apenas recursos disponíveis gratuitamente, desde o Sistema Operacional, IDE para a codificação, diversas API's e ferramentas para mineração e representação gráfica/estatística como Linguagem R e *Weka*.

Como proposta para possíveis trabalhos futuros, uma possibilidade seria utilizar as coordenadas geográficas oferecidas pelo *Twitter*, para desenvolver um gráfico que possibilite observar e comparar a polaridade das opiniões dos usuários em relação a um termo buscado em um mapa, facilitando a visualização das opiniões positivas e negativas dos usuários por regiões.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] NAVEGA, Sergio. Princípios Essenciais do Data Mining. InfoImagem, São Paulo, 2002.

[2] Fayyad, U. M., Shapiro, G. P., Uthurusamy, R. Advances in Knowledge Discovery and Data Mining. Cambridge, 1996.

[3] OLSON, D. L; DELEN, D. Advanced Data Mining Techniques. Springer, 2008.

[4] WITTEN, I. H; FRANK, E. Data Mining - Practical Machine Learning Tools and Techniques. Elsevier, 2005.

[5] HAN, Jiawei; KAMBER, Micheline. Data Mining: Concepts and Techniques. 2ª Edição. Urbana: Morgan Kaufmann, 2006.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009.

[7] GONÇALVES, L. P. F. Avaliação de ferramentas de mineração de dados como fonte de dados relevantes para a tomada de decisão. Dissertação (Mestrado em Administração), Universidade Federal do Rio Grande do Sul, 2004.

[8] TAURION, Cezar. Big data. Rio de Janeiro: Brasport, 2013.

[9] Descoberta do conhecimento (KDD). Disponível: <https://sites.google.com/site/mineracaodedados1b/descoberta-do- conhecimento-kdd>. Acesso em 21 maio, 2014.

[10] WEKA – Data Mining with Open Source Machine Learning Software in Java. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/index.html>. Acesso em 21 maio, 2014.

[11] R-Project - The R Projetect for Statistical Computing. Disnponível em: <http://www.r-project.org/>. Acesso em 21 maio, 2014.

[12] Free Statitics - Free Statical Software. Disponível em: <http://freestatistics.info/en/index.php>. Acesso em 21 maio, 2014.